DATA COLLECTION

W. T. Federer

BU-694-M

September 1979

Abstract

Concepts and definitions of various terms involved in collecting data are given. The various aspects of collecting data are discussed; these involve the why, what, how, where, who, and when aspects, as well as description, storage, and disposal of data. Three main types of data collection are in observational studies, in surveys and censuses, and in experimental investigations. A limited discussion of each is presented. Finally, some data collection agencies are listed, with most of those listed being in the United States.

Biometrics Unit Mimeo Series, Cornell University.

DATA COLLECTION

Walter T. Federer
Cornell University

1. Introduction

A distinction needs to be made between numbers, adjectives, and other forms of description and data. Numbers and adjectives can be available simply as entities in themselves or as data. The existence of numbers and adjectives does not imply the existence of a set of data, but the existence of a set of data implies the existence of some descriptive forms such as numbers, adjectives, phrases, pictures, graphs, etc. For example, the set of numbers {3,1,0,4,9,6} and the set of adjectives {small, pretty, personable, harsh, susceptible, resistant} do not in themselves convey information about any phenomenon. When numbers and adjectives convey information about some entity such as 0, 1, 3, and 4 worms in four particular apples and small, pretty, heat-resistant spores of a bacterium, these facts are denoted as data. A datum is defined to be a fact (numerical or otherwise) from which a conclusion may be drawn such as, for example, none of the four apples have the same number of worms. A datum contains information whereas a number, adjective, or other form of description may not.

The information in a set of data may be contaminated (mixed-up or confounded) with other kinds of information.

For example, the particular four apples may have been the only ones in a basket of apples which contained worms and the apples were from a part of the orchard which had a very light infestation of the codling moth. Thus, we do not know if four wormy apples per basket represents the proportion of wormy apples in the entire population of apples or only for this variety in lightly infested orchards. The various factors affecting variation and response are denoted as sources of variation in the data. When the various effects in a set of data are inseparable, they are said to be completely confounded. Partial confounding of effects occurs when the effects are partly mixed together, but it is possible to obtain estimates of each of the various sources of variation. No confounding of effects (orthogonality) implies maximum separability of sources of variation. For example, suppose that there is technician-to-technician variation, day-to-day variation, and method-to-method variation. If technician one does method one on day one, technician two does method two on day two, technician three does method three on day three, etc., one cannot separate the technician effect from the method effect or from the day effect. If, on the other hand, all technicians did all methods on each of several days, the different effects would be completely separable. The preceding plan would be an orthogonal (completely unconfounded) one, whereas the former would be completely confounded. Thus, the plan or

the design of the investigation has a large influence on the confounding aspects in data and hence on the amount of information derivable from the data. With adequate planning and foresight, data can be obtained which have little or no confounding of sources of variation of interest to the investigator. On the other hand, unplanned and/or haphazard collection of data, no matter how voluminous, most frequently results in highly confounded data sets which may be of little or no use in studying effects of various sources of variation. There is a tendency for people to believe that largeness of a data set removes the effects of confounding. This is not necessarily true. The closer one gets to obtaining responses for 100% of the population, the nearer one is to the values of the population parameters. In this sense, largeness does remove the difficulties of haphazard or selective sampling. However, the complete enumeration of a subpopulation provides no information on the other subpopulation parameters. In this sense, largeness of a set of data does nothing to remove the biases.

2. Aspects of Data Collection

Whenever data are to be collected, several aspects require consideration. The first one to consider is why should these data be collected and what uses will these data serve. If there is no purpose or no use for these data, why collect them? The ready accessibility of micro-computers, mini-

computers, and macro-computers makes it a simple matter to collect voluminous sets of data with relatively little effort. To illustrate, suppose that temperature in one-hundreth of a degree centigrade and humidity in one-hundreth of a percent are to be collected every second over a 20-year period in 1000 locations of a field. This results in 60 x 60 x 24 x 365.25 x 1000 x 20 = 631,150,000,000 measurements on temperature and the same number on humidity. The data set would consist of more than 1.25 trillion observations. Before embarking on such a large data collection venture, one should seriously consider taking fewer measurements (e.g., one every hour rather than 3600 per hour), the uses for data of this nature, and what individuals would actually use these data. Unfortunately, several individuals have and are collecting large data sets merely on the premise that they might be useful to someone at some time. They have a computer available and set about collecting data to make use of the computer.

A second aspect of data collection is <u>what</u> data will be collected. All pertinent data for studying a phenomenon should be collected. Thus it is essential to determine what data to collect in light of why the data are collected. Sufficient data should be collected to achieve the goals and purposes of the investigation. Provisions should also be made to obtain pertinent data that surface during the course of an investigation. The investigator must be aware of

evidence that comes to his attention, and obtain the necessary data to explain the phenomenon. An example was given above illustrating how a large data set could be obtained. In some investigations, such as sending missiles to the moon, one can obtain only a few observations on some entities. In both of these cases, it is essential to determine what data to collect.

A third aspect of data collection is how and where the data are to be collected. The statistician can be very helpful in designing and planning the investigation and in determining how and where the data are to be collected. The how and where of data collection are intimately entwined with the plan and type of the investigation. Also involved are how to measure and quantify information on the various phenomena in an investigation. This is especially true of social phenomenon. Methods of measuring responses must be carefully considered, and it should be ascertained if a measuring instrument is measuring what it is supposed to, or if it is actually measuring something else.

Two other aspects to be considered are who is to collect the data and when it is to be done. Unless these aspects are firmly regulated, the data may not be collected or only a portion of the data may be obtained. In many investigations it is imperative to have highly trained and unbiased technicians in order to carry out the investigation. The timing of an investigation may be important to its success, and/or

it may be necessary to carry out the investigation in a specified time period. If these aspects are ignored, the data may be so unreliable as to make them useless.

In addition to the why, what, how, where, and when aspects of data collection, it is imperative to have a complete, written description of all data obtained. This information recorded only in the investigator's mind may soon be lost, and may be unavailable to other investigators desiring to use the data. Plans need to be made for storage and/or disposal of all data collected. Discarding valuable data which are needed in future research or the storage of useless data result in economic losses which can be avoided with proper planning. Data may be stored in notebooks, on tapes, or other forms, but percautions need to be made to avoid loss or damage due to fire, water, tape erasures, etc.

3. Types of Data Collection

Three main types of data collection are observational investigations or studies, sample surveys and censuses, and experimental investigations. In the observational types of investigation, records are kept on whatever observations are available with little or no thought of making them representative of the population. In censuses, an attempt is made to obtain observations on every member of the population, whereas in sample surveys only a portion of the population is surveyed with an idea of the sample being representative of

the population. In an experiment conditions are often introduced which do not appear in the population, and there is a considerable degree of control over the conditions of the experiment. The controlled conditions in an experiment may not be available in observational and survey investigations.

3.1. Observational investigations. One of the large sources of observational data sets is record keeping. In all societies of the world, records are kept on a wide variety of phenomena. For example, there are records of births, of deaths, of marriages, of church members, of traffic tickets issued, of convictions, of diplomas, of fraternal and social organizations, of gem collectors, of daily maximum and minimum temperatures, of daily rainfall, of auto sales by dealerships, of treatments administered to patients by doctors, and on and on. In totality, records of events comprise a very voluminous set of data. These records are often useful for the purpose for which they were collected, but may be useless for another purpose because of the method for determining whether or not a record would be kept. For example, in the issuance of traffic tickets, no record may be kept of traffic tickets that were issued and then destroyed. Thus, the records available on traffic tickets issued may be only those tickets issued which are eventually presented to a judge with no mention being made of the number of tickets issued and destroyed before presentation. The omission of a segment of the

population would make the remaining data useless in determining ratio of convictions to number of traffic tickets issued. One of the major faults of observational data is the omission of subsets of the data without knowing or having a description of the subsets omitted. In many cases, no valid conclusions are possible about the entire population from a portion of the subpopulations in the observational data set.

Observational data sets are often used in studies simply because of their availability. Some statistical studies have been made for utilizing observational data, but no general methods exist for drawing valid conclusions from observational data sets. Each set has to be considered on its own merits, and the investigator must exercise caution in making inferences from the data to the population.

3.2. Censuses and sample surveys. The population to be surveyed is defined and then either a 100% sample (a census) or a smaller percentage of the population is surveyed. Interest centers on the conditions and events that occur within the specified population. Censuses are usually not 100% samples because some sampling units are inaccessible and others may appear more than once. However, an attempt is made to obtain a 100% sample with no repetitions. For example, the U. S. Department of Commerce, through its Census Bureau, is bound by law to carry out a population census for the entire United States every ten years. This is an enormous and costly task

which has omissions and duplications despite elaborate precautions. It is an impossible task to take censuses in countries with very large populations as in India and China. Instead, these countries must obtain population estimates through sample survey methods. The sample percentage may be small, say 0.1% to 1%, in order to obtain sufficient trained personnel to conduct the survey.

Sample surveys may be grouped into two broad categories: probability sample surveys and nonprobability sample surveys. In the former type, the probability of selecting sampling units from the population is known, whereas it is not in the latter type. Whether or not nonprobability sample estimates can be regarded as representative of the population parameters is unknown. Despite this, a large number of surveys are of the nonprobability type. The lower cost and convenience of such surveys is attractive to investigators even if they are unable to check on the biases in the survey.

Some types of nonprobability sample surveys are:

i) Judgement – The investigator limits the selection of the sampling units to those he judges to be representative of the population.

ii) Convenience – The investigator selects the sampling units which are convenient or readily accessible.

iii) Quota – The only restriction on the selection of the sampling units if that there be a specified number in each

of the specified groups. A quota in each group is thus maintained by a variety of unspecified sampling procedures.

iv) Purposely biased – The investigator devises a sampling procedure which eliminates all sampling units of an undesirable class or he selects only the sampling units which give the desired result, say two to one, nine out of ten, etc.

v) Haphazard – The investigator selects the sampling units in such a manner as to leave the reader with the impression that the selection involved randomization and a probability survey.

Some types of probability sample surveys are:

i) Simple random sample – The population does not contain subpopulations and each sampling unit in the population has an equal and independent chance of being selected. An equivalent definition is that every possible sample of size n has an equal chance of being selected.

ii) Stratified-simple random sample – The population is composied of subpopulations and a simple random sample of sampling units is selected within <u>each</u> of the subpopulations (strata).

iii) Cluster-simple random sample – The population is composed of subpopulations (clusters) but a simple random sample of subpopulations (clusters) is made; then, a simple random sample is made within each of the selected clusters. When the clusters are areas, the sample survey design is known as

an area-simple random sample.

iv) Every k^{th} sampling unit with a random start — In many cases, a list of sampling units or a serial ordering of sampling units is available. The list is partitioned in subgroups of k items each. A number between 1 and k, say t, is randomly selected. Then, the t^{th} item, the $(t+k)^{th}$ item, the $(t+2k)^{th}$ item, the $(t+3k)^{th}$ item, etc. forms the sample. This survey design has also been denoted as systemmatic sampling.

v) More complex designs — There are many types of sample survey designs involving more complexity than the above. Several of the above types may be included in these designs.

Probability sample survey designs are recommended for use in data collection because their properties are known. Representative and unbiased estimates of population parameters are possible with these designs, but not with the others described above.

3.3. Experiment designs. Every experiment involves data collection and has a plan of procedure, some involving randomization and some not. Because of the statistical properties of the randomized designs, they are recommended over systemmatic or nonrandomized designs. In order to have higher precision (repeatability), blocking is used to group the units (experimental units) used in the experiment into blocks, or groups, which are relatively homogeneous within blocks. The blocks

are the subpopulation (clusters) in the population. A simple random sample of blocks is made and a simple random sample of units in each block is made. The entities of interest are called treatments. These are randomly allocated to the experimental units within each block in the randomized design and selectively placed in the systemmatic ones. Then, the experiment is put into operation and data are collected on each experimental unit in order to obtain information as a basis for comparing treatments and/or of describing the action of treatments on the responses for each experimental unit, and responses may be obtained for a large number of variables.

Experiment designs (plans for arrangement of treatments in an experiment) are used for data collection in research, developmental, and investigational studies. They are used under relatively controlled conditions as compared to survey designs. A selected set of phenomena are studied in any particular experiment and these may or may not be present in the general population.

Some types of experiment design are:

i) Unblocked – The completely randomized block design is a member of this class. The treatments are randomly allocated to the experimental units in the experiment and there is no blocking.

ii) One-way blocking – The randomized complete and incomplete block designs are members of this class. Complete block designs

have all treatments in each block, whereas various subsets of treatments are used in the blocks for an incomplete block design. The latter are used when the block sizes are too small to accommodate all treatments in a block. There are numerous classes of incomplete block designs.

iii) Two-way blocking – In some experimental situations, variation exists in two directions or for two sources of experimental variation. In order to remove these sources of variation from the comparisons between treatment effects, it is necessary to use two-way blocked designs. Some examples are latin square, Youden, F-square, change-over, and other row by column designs.

iv) Three-way and higher way blocking.

4. Data Collection Agencies

Data collection agencies appear in every part of a community. The judicial, legislative, and executive branches of village, city, county, state, national, and international governments have a variety of data collection agencies. Some of the better known U. S. Federal data collection agencies are the Census Bureau, the Bureau of Labor Statistics, the Statistical Reporting Service, the Central Intelligence Agency, the Internal Revenue Service, the Food and Drug Administration, the Bureau of the Budget, the General Services Administration, etc. In addition to the agencies located in Washington, D. C., branches are located throughout the United States.

The United Nations has several large data collection agencies, as do all the governments of the world. The more developed a country becomes the more extensive and sophisticated are the data collection agencies. The data are often utilized in planning the economic, business, and social policies of a country. In order to have intelligent planning, it is necessary to have data on the phenomenon at hand.

In the private sector of the U. S., there are many national and state survey organizations collecting data for specific purposes. The Roper, Gallup, and Harris survey organizations provide the polls frequently reported in newspapers. In addition, volumes of data are collected by pharmaceutical, industrial, business and educational organizations. The Wall Street Journal is devoted to reporting data on business transactions on stocks, bonds, mutuals, futures, etc. Large quantities of data are accummulated and reported every week-day.

Research organizations of colleges, universities, industries, corporations, institutes, etc., conduct and report a vast number of research, development, and investigational studies and surveys. Data from these studies are reported in scientific journals of which there is a very large number. The reporting of results from these studies is an effort to make the results of data collections and the resulting conclusions available to the general scientific population.

From the above it should be apparent that a vast amount of data collection is made in all countries of the world and that individual, corporate, and political relations are affected to a considerable degree by the data collected. Data collection is an inescapable and essential part of the lives of all individuals in a community. Hence, there should be considerable desire to obtain the best and most complete data possible on the desired variables of interest. Statisticians can be of considerable aid in the planning of all data collection programs.

5. Further Reading

Federer, W. T. (1973). Statistics and Society – Data Collection and Interpretation. Marcel Dekker, Inc., New York, Chapters I to VI, XIII.

Wallis, W. A., et al. (1971). Federal Statistics – A Report of the President's Commission, volumes I and II. Supt. of Documents, U. S. Government Printing Office, Washington, D. C.